

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-245068

(43)Date of publication of application : 30.08.2002

(51)Int.Cl.

G06F 17/30

G06F 3/14

G06F 13/00

(21)Application number : 2001-034718

(71)Applicant : INTERNATL BUSINESS MACH
CORP <IBM>

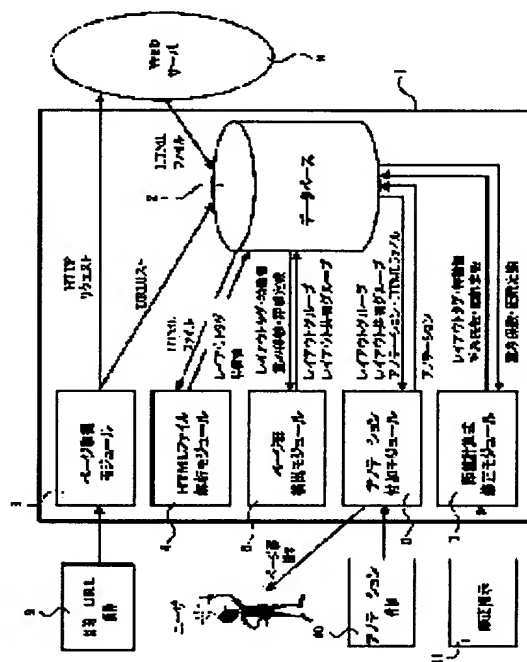
(22)Date of filing : 09.02.2001

(72)Inventor : FUKUDA KENTARO
TAKAGI HIRONOBU
MAEDA JUNJI
ASAKAWA CHIEKO(54) METHOD AND SYSTEM FOR PROCESSING INFORMATION, PROGRAM AND
RECORDING MEDIUM

(57)Abstract:

PROBLEM TO BE SOLVED: To improve efficiency in a work for imparting an annotation to an HTML page file.

SOLUTION: A page obtaining module 3 obtains the page file from a web server 8 and an HTML file analyzing module 4 derives a tag and featured values concerning a layout. A page group detecting module 5 makes the page file where the layout is the same or similar into a group through the use of the layout tag and the featured values. When the annotation is added to the optional page file of the grouped layout group by using an annotation adding module 6, the annotation is applicable concerning another page file in the layout group. When a user divides or synthesizes the layout group, a distance calculation formula between the pages or the layout groups is corrected by a distance calculation formula correcting module 7 and the result of division or synthesis by the user is reflected.



LEGAL STATUS

[Date of request for examination]

16.10.2001

[Date of sending the examiner's decision of
rejection][Kind of final disposal of application other than
the examiner's decision of rejection or
application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision
of rejection]

[Date of requesting appeal against examiner's
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開2002-245068

(P2002-245068A)

(43)公開日 平成14年8月30日(2002.8.30)

(51)Int.Cl. ⁷	識別記号	F I	テーマコード [*] (参考)
G 0 6 F 17/30	2 2 0	G 0 6 F 17/30	2 2 0 A 5 B 0 6 9
	1 1 0		1 1 0 F 5 B 0 7 5
	2 1 0		2 1 0 D
3/14	3 1 0	3/14	3 1 0 C
13/00	5 5 0	13/00	5 5 0 L
審査請求 有 請求項の数22 O L (全 19 頁)			

(21)出願番号 特願2001-34718(P2001-34718)

(22)出願日 平成13年2月9日(2001.2.9)

(71)出願人 390009531

インターナショナル・ビジネス・マシー
ズ・コーポレーション

INTERNATIONAL BUSIN
ESS MASCHINES CORPO
RATION

アメリカ合衆国10504、ニューヨーク州

アーモンク ニュー オーチャード ロー
ド

(74)復代理人 100112520

弁理士 林 茂則 (外3名)

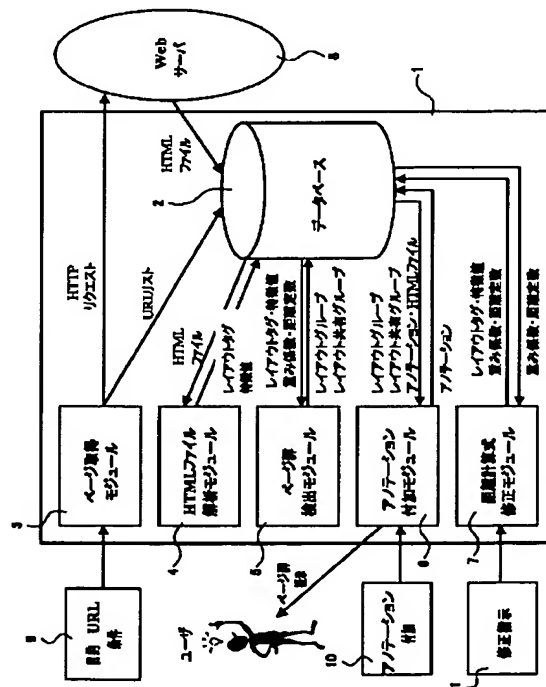
最終頁に続く

(54)【発明の名称】 情報処理方法、情報処理システム、プログラムおよび記録媒体

(57)【要約】

【課題】 HTMLページファイルへのアノテーション付与の作業効率を向上する。

【解決手段】 ページ取得モジュール3でウェブサーバ8からページファイルを取得し、HTMLファイル解析モジュール4でレイアウトに関するタグと特徴値を導出する。レイアウトタグとその特徴値からレイアウトが同一または類似するページファイルをページ群検出モジュール5を用いてグループ化する。グループ化されたレイアウトグループの任意のページファイルにアノテーション付加モジュール6を用いてアノテーションを付加すると、レイアウトグループ内の他のページファイルについてもそのアノテーションが適用される。レイアウトグループがユーザによって分割されたり、統合された時には、距離計算式修正モジュール7によってページ間あるいはレイアウトグループ間の距離計算式が修正され、ユーザによる分割あるいは統合の結果を反映する。



【特許請求の範囲】

【請求項 1】 ウェブサイトのページファイルを取得するステップと、
前記ページファイルのページレイアウト構造が互いに同一または類似する前記ページファイルのグループを生成するステップと、
前記グループ内の任意のページファイルにアノテーションを付与するステップと、
前記グループ内の他のページファイルの全部または一部に前記アノテーションに関連付けるステップと、
を含む複数のページファイルにアノテーションを付加するための情報処理方法。

【請求項 2】 前記グループを生成するステップには、
前記ページレイアウト構造の構造記述式と前記構造記述式の特徴値とを前記ページファイルの解析によって導出するステップと、
前記構造記述式および特徴値を用いて、前記ページファイル間の類似性を表すページ間距離を算出するステップと、
前記ページ間距離が所定の値以下の範囲の前記ページファイルをグループ化するステップと、
を含む請求項 1 記載の情報処理方法。

【請求項 3】 前記構造記述式は、前記ページファイルに含まれる前記ページレイアウト構造に関連するタグを、ページ内での位置を特定するための書式を用いて表現したレイアウトタグであり、
前記特徴値は、前記レイアウトタグの属性および前記属性の値である請求項 2 記載の情報処理方法。

【請求項 4】 前記ページ間距離は、複数の前記ページファイルに共通に含まれる前記構造記述式について、前記構造記述式と前記特徴値とを重み付けした値の和をとることにより算出する請求項 2 記載の情報処理方法。

【請求項 5】 前記アノテーションを前記グループの他のページファイルに関連付けるステップには、
前記アノテーションを前記グループ内の全てのページファイルに適用するかを判断するステップと、
前記判断が偽の場合に、前記アノテーションが関連付けられないページファイルで構成されるページ群の任意のページファイルに第 2 のアノテーションを付加するステップと、
前記ページ群内の他のページファイルの全部または一部に前記第 2 のアノテーションに関連付けるステップと、
前記アノテーションが関連付けられたページファイルと、前記第 2 のアノテーションが関連付けられたページファイルとが、前記グループを生成するステップにおいて同一グループにグループ化されないように、前記ページ間距離の計算式を修正するステップと、
を含む請求項 2 記載の情報処理方法。

【請求項 6】 前記ページ間距離は、複数の前記ページファイルに共通に含まれる前記構造記述式について、前

10

20

30

40

50

記構造記述式と前記特徴値とを重み付けした値の和をとることにより算出され、
前記ページ間距離の計算式の修正は、
前記アノテーションが関連付けられたページファイルと前記第 2 のアノテーションが関連付けられたページファイルとの間で相違する前記構造記述式または特徴値について、その構造記述式および特徴値の前記重みを増加する操作、または、
前記アノテーションが関連付けられたページファイルと前記第 2 のアノテーションが関連付けられたページファイルとの間で共通する前記構造記述式または特徴値について、その構造記述式および特徴値の前記重みを減少する操作、
の何れかの操作により行う請求項 5 記載の情報処理方法。

【請求項 7】 前記グループを代表する代表構造記述式と、前記代表構造記述式の代表特性値とを導出するステップと、
前記代表構造記述式および代表特徴値を用いて、前記グループ間の類似性を表すグループ間距離を算出するステップと、
前記グループ間距離が所定の値以下の範囲の前記グループに含まれる前記ページファイルをグループ化し共有グループを生成するステップと、
前記共有グループに含まれる任意のページファイルにおいて、そのページレイアウト構造の一部が、他のページファイルのページレイアウト構造の一部または全部と一致または類似する共有領域にアノテーションを付加するステップと、
前記共有グループに含まれる他のページファイルの前記共有領域に前記アノテーションに関連付けるステップと、
をさらに含む請求項 2 記載の情報処理方法。

【請求項 8】 前記代表構造記述式は、前記ページファイルに含まれる前記ページレイアウト構造に関連するタグを、ページ内での位置を特定するための書式を用いて表現したレイアウトタグであり、
前記代表特徴値は、前記レイアウトタグの属性および前記属性の値である請求項 7 記載の情報処理方法。

【請求項 9】 前記グループ間距離は、複数の前記グループに共通に含まれる前記代表構造記述式について、前記代表構造記述式と前記代表特徴値とを重み付けした値の和をとることにより算出する請求項 7 記載の情報処理方法。

【請求項 10】 前記アノテーションを前記他のページファイルの共有領域に関連付けるステップには、
前記アノテーションを前記共有グループ内の全てのページファイルの共有領域に適用するかを判断するステップと、
前記判断が偽の場合に、前記アノテーションが関連付け

られない共有領域を含むページファイルで構成されるページ群の任意のページファイルの共有領域に第 2 のアノテーションを付加するステップと、
前記ページ群内の他のページファイルの共有領域の全部または一部に前記第 2 のアノテーションを関連付けるステップと、
前記アノテーションが関連付けられた共有領域を含むページファイルと、前記第 2 のアノテーションが関連付けられた共有領域を含むページファイルとが、前記共有グループを生成するステップにおいて同一共有グループにグループ化されないように、前記グループ間距離の計算式を修正するステップと、
を含む請求項 7 記載の情報処理方法。

【請求項 1 1】 ウェブサイトのページファイルを取得する手段と、
前記ページファイルのページレイアウト構造が互いに同一または類似する前記ページファイルのグループを生成する手段と、
前記グループ内の任意のページファイルにアノテーションを付与する手段と、前記グループ内の他のページファイルの全部または一部に前記アノテーションを関連付ける手段と、
を含む複数のページファイルにアノテーションを付加するための情報処理システム。

【請求項 1 2】 前記グループを生成する手段には、前記ページレイアウト構造の構造記述式と前記構造記述式の特徴値とを前記ページファイルの解析によって導出する手段と、
前記構造記述式および特徴値を用いて、前記ページファイル間の類似性を表すページ間距離を算出する手段と、
前記ページ間距離が所定の値以下の範囲の前記ページファイルをグループ化する手段と、
を含む請求項 1 1 記載の情報処理システム。

【請求項 1 3】 前記構造記述式は、前記ページファイルに含まれる前記ページレイアウト構造に関連するタグを、ページ内での位置を特定するための書式を用いて表現したレイアウトタグであり、
前記特徴値は、前記レイアウトタグの属性および前記属性の値である請求項 1 2 記載の情報処理システム。

【請求項 1 4】 前記ページ間距離は、複数の前記ページファイルに共通に含まれる前記構造記述式について、前記構造記述式と前記特徴値とを重み付けした値の和をとることにより算出する請求項 1 2 記載の情報処理システム。

【請求項 1 5】 前記アノテーションを前記グループの他のページファイルに関連付ける手段には、
前記アノテーションを前記グループ内の全てのページファイルに適用するかを判断する手段と、
前記判断が偽の場合に、前記アノテーションが関連付けられないページファイルで構成されるページ群の任意の

ページファイルに第 2 のアノテーションを付加する手段と、
前記ページ群内の他のページファイルの全部または一部に前記第 2 のアノテーションを関連付ける手段と、
前記アノテーションが関連付けられたページファイルと、前記第 2 のアノテーションが関連付けられたページファイルとが、前記グループを生成する手段において同一グループにグループ化されないように、前記ページ間距離の計算式を修正する手段と、
を含む請求項 1 2 記載の情報処理システム。

【請求項 1 6】 前記ページ間距離は、複数の前記ページファイルに共通に含まれる前記構造記述式について、前記構造記述式と前記特徴値とを重み付けした値の和をとることにより算出され、
前記ページ間距離の計算式の修正は、
前記アノテーションが関連付けられたページファイルと前記第 2 のアノテーションが関連付けられたページファイルとの間で相違する前記構造記述式または特徴値について、その構造記述式および特徴値の前記重みを増加する操作、または、
前記アノテーションが関連付けられたページファイルと前記第 2 のアノテーションが関連付けられたページファイルとの間で共通する前記構造記述式または特徴値について、その構造記述式および特徴値の前記重みを減少する操作、
の何れかの操作により行われる請求項 1 5 記載の情報処理システム。

【請求項 1 7】 前記グループを代表する代表構造記述式と、前記代表構造記述式の代表特性値とを導出する手段と、
前記代表構造記述式および代表特徴値を用いて、前記グループ間の類似性を表すグループ間距離を算出する手段と、
前記グループ間距離が所定の値以下の範囲の前記グループに含まれる前記ページファイルをグループ化し共有グループを生成する手段と、
前記共有グループに含まれる任意のページファイルにおいて、そのページレイアウト構造の一部が、他のページファイルのページレイアウト構造の一部または全部と一致または類似する共有領域にアノテーションを付加する手段と、
前記共有グループに含まれる他のページファイルの前記共有領域に前記アノテーションを関連付ける手段と、
をさらに含む請求項 1 2 記載の情報処理システム。

【請求項 1 8】 前記代表構造記述式は、前記ページファイルに含まれる前記ページレイアウト構造に関連するタグを、ページ内での位置を特定するための書式を用いて表現したレイアウトタグであり、
前記代表特徴値は、前記レイアウトタグの属性および前記属性の値である請求項 1 7 記載の情報処理システム。

【請求項 19】 前記グループ間距離は、複数の前記グループに共通に含まれる前記代表構造記述式について、前記代表構造記述式と前記代表特徴値とを重み付けした値の和をとることにより算出する請求項 17 記載の情報処理システム。

【請求項 20】 前記アノテーションを前記他のページファイルの共有領域に関連付ける手段には、前記アノテーションを前記共有グループ内の全てのページファイルの共有領域に適用するかを判断する手段と、前記判断が偽の場合に、前記アノテーションが関連付けられない共有領域を含むページファイルで構成されるページ群の任意のページファイルの共有領域に第 2 のアノテーションを付加する手段と、前記ページ群内の他のページファイルの共有領域の全部または一部に前記第 2 のアノテーションに関連付ける手段と、前記アノテーションが関連付けられた共有領域を含むページファイルと、前記第 2 のアノテーションが関連付けられた共有領域を含むページファイルとが、前記共有グループを生成する手段において同一共有グループにグループ化されないように、前記グループ間距離の計算式を修正する手段と、を含む請求項 17 記載の情報処理システム。

【請求項 21】 複数のページファイルにアノテーションを付加するためのコンピュータプログラムであって、コンピュータに、ウェブサイトのページファイルを取得する機能と、前記ページファイルのページレイアウト構造が互いに同一または類似する前記ページファイルのグループを生成する機能と、前記グループ内の任意のページファイルにアノテーションを付与する操作を要求する機能と、前記グループ内の他のページファイルの全部または一部に前記アノテーションに関連付ける機能と、を実現させるためのプログラム。

【請求項 22】 複数のページファイルにアノテーションを付加するためのコンピュータプログラムが記録されたコンピュータ可読な記録媒体であって、コンピュータに、ウェブサイトのページファイルを取得する機能と、前記ページファイルのページレイアウト構造が互いに同一または類似する前記ページファイルのグループを生成する機能と、前記グループ内の任意のページファイルにアノテーションを付与する操作を要求する機能と、前記グループ内の他のページファイルの全部または一部に前記アノテーションに関連付ける機能と、を実現させるプログラムが記録された記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、情報処理方法および情報処理システムに関する。特に、一般的なウェブページ用に作成されたページデザインを小面積デバイス等に表示するための変換ソフトや音声ブラウザ等の読み上げソフトに利用できるアノテーションの付与効率の向上に適用して有効な技術に関する。

【0002】

【従来の技術】 インターネット利用の一般化および多様化を背景に、インターネットに接続するための機器も多様化している。すなわち、従来インターネットに接続するための機器として、対角 12 インチ乃至 20 インチ程度の表示面積を有する CRT (Cathode Ray Tube)、液晶表示装置、プラズマディスプレイ装置等を備えたコンピュータシステムが利用されている。

【0003】 しかし、携帯性を重視する場合や携帯電話とのインテグレーションを考慮して PDA (Personal digital assistants) や i モード携帯電話の普及が急速に広がりつつある。これら携帯性を重視する機器では、表示面積が小さいのが一般的である。また、たとえば視覚障害者の場合、表示装置を視認してコンピュータ出力を確認できないので、たとえば音声ブラウザ等の読み上げソフトの利用が図られる。これら読み上げソフトは視覚障害者のみならず、コンピュータに慣れ親しまない人々とのヒューマンインタフェースを向上することも期待され、コンピュータシステムのより広い普及を促す技術として期待される。さらに、ウェアラブルコンピュータの場合には、必然的に表示装置の面積は小さくならざるを得ないので、音声出力が主または補助的な出力手段になると予想される。

【0004】 ところで、一般的なウェブサイトにおけるページレイアウトは、対角 12 インチ乃至 20 インチ程度の表示面積を有する通常のコンピュータシステムの表示装置を前提にデザインされる。また、表示装置への出力を前提にしているので、もとより視認できる暗眼者の使用を前提にしている。つまり表示部の上部あるいは左側に、サイト内のメニュー領域 (リンク情報等が埋め込まれる) や広告用のバナー等が配置され、暗眼者が 2 次的に視認しやすいように各情報の表示位置がレイアウトされる。通常そのページ固有の情報はページレイアウトの中央部や後半部分に配置されることが多い。

【0005】 このような暗眼者用かつ大面積の表示装置用にレイアウトされているウェブページを PDA や携帯電話あるいは音声ブラウザで表示または出力しようとした場合、通常ページの先頭部分に配置される情報 (フレーム情報や広告等) が邪魔になる場合が多い。これらフレーム情報や広告等の二次的な情報は、暗眼者用大面積表示においては表示を見やすくし、また使い勝手を向上するために有用であるが、小面積デバイスや音声ブラウザのユーザのためには、本来必要なそのページに固有の情報こそが必要であり、前記二次的な情報はむしろ邪魔に

なる。

【0006】そこで、大面積用にデザインされたページファイルを小画面デバイスや音声ブラウザを用いて出力する場合、本来必要な固有の情報に素早くアクセスする手段が必要である。この手段の一例に、ページファイルにアノテーションを付与する手法が知られている。すなわち、ページファイルの構造や各部位の重要度等の情報を外部ファイルに記述し、ページファイルを簡略化する際にこのアノテーションを参照してより迅速にまた高精度に簡略化を実行できるようにする。

【0007】

【発明が解決しようとする課題】しかし、各ページファイルへのアノテーションの付与は容易ではない。一般に各ページファイルをブラウジングし、表示を確認しながら重要度を判断してアノテーションを付す作業が必要になり、ボランティア等の人手に頼らざるを得ない。特に、ニュースサイトやデータベースサイトの場合にはページファイルが大量に存在するので、アノテーション付与の労力が膨大になる。また、たとえばニュースファイルのようにページファイルのURL (Uniform Resource 20 Locator) に日付データを含めて煩雑に新規ファイルが生成されるような場合には、一旦アノテーションを付しても再度アノテーション付与を行う必要がある。

【0008】本発明の目的は、ページファイルへのアノテーション付与の作業を効率的に行う手法を提供することにある。

【0009】

【課題を解決するための手段】本発明の概要を説明すれば以下の通りである。すなわち、本発明は、HTML (Hypertext Markup Language) 文書等におけるタグの 30 構造から同一のレイアウトを用いたページ群を検出し、それらのページ間でアノテーションを共有するための手法である。ユーザが指定したサイト内のページに対しHTML文書等レイアウト構造を有するファイルの解析を行い、レイアウトを決定する要因となるタグ (以下レイアウトタグと称する) を列挙する。この際、HTML等の文書内でのレイアウトタグの構造が明らかになるよう、ページ内での位置を特定するための書式、たとえばXPath、Xpointerやツリー形式などの書式を用いた構造記述式となるようレイアウトタグを記述するとともに、それぞれのレイアウトタグ (構造記述式) の特徴値を導出する。次にそれらの情報を基に、ページ間の距離を計算することにより、同一レイアウトを用いているページ群およびレイアウトの一部を共有しているページ群を自動的に検出しユーザに提示する。そして、ユーザにより、ページ群を代表する1つのページにアノテーションが付加されると、同一レイアウトを用いているページ群の全てに対応するアノテーションを付加する。また、レイアウトを共有しているページがあった場合は、まず、共有部分へのアノテーション付加を行い、 50

次にそれぞれのページ群が独立に保持している部分へのアノテーション付加を行う。これにより効率的なアノテーション付加が可能となる。

【0010】また、本発明では、ユーザが提示されたページ群をさらに分割・統合するなどの修正を加えた場合には、その結果を用いて距離計算式を修正することができる。これにより、以降のページ群分割の精度を向上させることができる。

【0011】

10 【発明の実施の形態】以下、本発明の実施の形態を図面に基いて詳細に説明する。ただし、本発明は多くの異なる態様で実施することが可能であり、本実施の形態の記載内容に限定して解釈すべきではない。なお、実施の形態の全体を通して同じ要素には同じ番号を付するものとする。

【0012】以下の実施の形態では、主に方法またはシステムについて説明するが、当業者であれば明らかとなり、本発明は方法、システムその他、コンピュータで使用可能なプログラム、あるいはプログラムが記録された媒体としても実施できる。したがって、本発明は、ハードウェアとしての実施形態、ソフトウェアとしての実施形態またはソフトウェアとハードウェアとの組合せの実施形態をとることができる。プログラムが記録された媒体としては、ハードディスク、CD-ROM、光記憶装置または磁気記憶装置を含む任意のコンピュータ可読媒体を例示できる。

【0013】また以下の実施の形態では、一般的なコンピュータシステムを用いることができる。実施の形態で用いるコンピュータシステムには、中央演算処理装置 (CPU)、主記憶装置 (メインメモリ: RAM)、不揮発性記憶装置 (ROM) 等を有し、バスで相互に接続される。バスには、その他コプロセッサ、画像アクセラレータ、キャッシュメモリ、入出力制御装置 (I/O) 等が接続されてもよい。バスには、適当なインタフェースを介して外部記憶装置、データ入力デバイス、表示デバイス、通信制御装置等が接続される。その他、一般的にコンピュータシステムに備えられるハードウェア資源を備えることが可能なことは言うまでもない。外部記憶装置は代表的にはハードディスク装置が例示できるが、これに限られず、光磁気記憶装置、光記憶装置、フラッシュメモリ等半導体記憶装置も含まれる。データ入力デバイスには、キーボード等の入力装置、マウス等ポインティングデバイス、ペン入力装置、タブレット装置等を備えることができる。データ入力デバイスにはスキャナ等の画像読み取り装置、音声入力装置も含む。表示装置としては、CRT、液晶表示装置、プラズマ表示装置が例示できる。また、コンピュータシステムには、パーソナルコンピュータ、ワークステーション、メインフレームコンピュータ等各種のコンピュータが含まれる。

【0014】図1は、本発明の一実施の形態である情報

処理システムの一例を示したブロック図である。本実施の形態の情報処理システム1は、データベース2、ページ取得モジュール3、HTMLファイル解析モジュール4、ページ群検出モジュール5、アノテーション付加モジュール6、距離計算式修正モジュール7を有する。

【0015】データベース2は、後述する各モジュールで生成されるデータやウェブサーバ8から取得したページファイル（HTMLファイルともいう）を記録する。データベース2は本実施の形態の情報処理システム1内に備えられるハードディスクドライブ等の記憶装置とデータの入出力を制御するソフトウェアで構成される。しかし必ずしも情報処理システム1内に備えられる必要はなく、たとえばURLによって特定される外部ファイルであっても良い。また、データベース2は集中的に管理される必要もなく、分散的に記録管理されてもよい。つまり、適当なアドレス指定手段によって必要なデータの入出力が実現できる限り、物理的な記録デバイスの存在場所や形態に関わらず、本実施の形態のデータベース2を構成できる。

【0016】ページ取得モジュール3は、ユーザによる目的URL条件9の入力を受けて、目的URLとそれに関連するURLの内容であるウェブページをウェブサーバ8から取得する。取得要求は、たとえばHTTP（Hypertext Transfer Protocol）により発する。取得された目的URLのHTMLファイル（ページファイル）はデータベース2に記録される。

【0017】ページ取得モジュール3は、まず、目的URLを含むURLリストを生成し、目的URLのページファイルを取得する。次に、目的URLのサイト内のページをリストアップすることにより関連するURLを取得する。たとえば、目的URLのページ内に含まれるURL（たとえば<a>タグのhrefアトリビュートから得る）を列挙する。この中からユーザの指定した範囲内に含まれるURLのみを選択し、URLリストに加える。次にURLリスト内のページを順に取得し、取得したページファイルをデータベース2に記録する。取得した関連URLに関連するURLを含む場合には、それぞれの関連URLに対して同様の操作を再帰的に行う。これにより、サイト内でリンクが張られたページを取得できる。この際、既にURLリストに出現したURLの二重登録は行わないようにする。URLリストはデータベース2に記録される。

【0018】HTMLファイル解析モジュール4は、ページ取得モジュール3で取得したページファイルを解析し、ページレイアウトに影響を及ぼすレイアウトタグのリストアップとその特徴値の導出とを行う。

【0019】図2は、HTMLファイル解析モジュール4の構成の一例を示したブロック図である。HTMLファイル解析モジュール4は、HTMLパーザ20とレイアウトタグリストアップモジュール21と特徴値導出モ

ジュール22とを有する。

【0020】HTMLパーザ20は、ページ取得モジュール3で取得されたHTMLファイル（ページファイル）を解析し、DOMツリー等、タグの構造記述式が容易に取得できる形式に変換する。

【0021】レイアウトタグリストアップモジュール21は、HTMLパーザ20で得られたタグ構造について、レイアウト構造に影響を与えるタグ（レイアウトタグ）を構造記述式を用いてリストアップする。なお、レイアウトタグとしては、例えば、「table」、「tbody」、「tr」、「td」、「th」、「hr」などが例示できる。また、構造記述式には、例えばXPath、XPathPointerなどのページ内での位置を特定するための書式あるいはツリー形式等の書式を適用できる。

【0022】特徴値導出モジュール22は、リストアップされたレイアウトタグについて、当該タグがもつアトリビュート（Attribute：属性ともいう）や、当該タグのサブツリー内に含まれるエレメント（element：要素ともいう）等の情報をその特徴値として構造記述式に関連付ける。特徴値として、以下のようなアトリビュートやエレメントを例示できる。すなわち、レイアウトタグ「table」に対しては、「align」、「bgcolor」、「border」、「cellpadding」、「cellspacing」、「width」が、レイアウトタグ「tbody」に対しては、「align」、「valign」が、レイアウトタグ「tr」に対しては、「align」、「bgcolor」、「valign」が、レイアウトタグ「td」あるいは「th」に対しては、「align」、「bgcolor」、「colspan」、「height」、「rowspan」、「valign」、「width」などのアトリビュート（属性）に加えテキストや画像などのエレメント（要素）の有無およびそのサイズなどが、レイアウトタグ「hr」に対しては、「align」、「width」、「size」、「noshade」が例示できる。

【0023】HTMLファイル解析モジュール4は、レイアウトタグリストアップモジュール21および特徴値導出モジュール22で取得され、構造記述式で表現されたレイアウトタグおよびそれに関連付けられた特徴値を前記URLリストの各URLに関連付け、データベース2に記録する。

【0024】図3は、URLリストに含まれるURLとそのURLに関連するレイアウトタグと特徴値の例を示した図である。たとえば「http://www.ibm.com/index.html」のURLには、構造記述式（この場合XPath）で記述されたレイアウトタグ「/html[1]/body[1]/table[1]」と、「/html[1]/body[1]/table[1]/tr[1]/td[1]」とが含まれ、「/html[1]/body[1]/table[1]」には、「w

11

idth=200,bgcolor=blue,...」の特徴値が関連付けられ、
「/html[1]/body[1]/table[1]/tr[1]/td[1]」には、「bgcolor=red,...」の特徴値が関連付けられている。

【0025】ページ群検出モジュール5は、HTMLファイル解析モジュール4によって導出されたレイアウトタグおよびその特徴値を利用してページ間距離を求める機能をもつ。この機能を用いて同一あるいは類似のレイアウト構造をもつページ群をレイアウトグループとして抽出する。また、ページファイルの一部の領域について、他のページファイルと共通のレイアウト構造をもつ共通レイアウトを算出し、それらをレイアウト共有グループとして抽出する機能を持つ。

【0026】図4は、ページ群検出モジュール5の構造の一例を示したブロック図である。ページ群検出モジュール5は、ページ間距離計算モジュール41、レイアウトグループ判別モジュール42、代表値計算モジュール43、レイアウト間距離計算モジュール44、レイアウト共有グループ判別モジュール45を有する。

【0027】ページ間距離計算モジュール41は、レイアウトタグに関連付けられている特徴値から、そのレイアウトタグが属するページファイルと他のページファイルとの距離を計算する機能を持つ。レイアウトグループ判別モジュール42は、ページ間距離計算モジュール41で計算されたページ間距離が互いに所定の範囲内にあるページファイルをレイアウトグループとして抽出する。代表値計算モジュール43は、レイアウトグループとして群化された互いに同一または類似のレイアウト構造を有するページファイル群の代表値を計算する。レイアウト間距離計算モジュール44は、レイアウトグループ間の距離を計算する機能を持ち、レイアウト共有グループ判別モジュール45は、レイアウトグループを構成するページファイルの一部の領域について他のレイアウトグループを構成するページファイルで共通に持つ同一または類似のレイアウト構造があるかを計算し、このような共有レイアウトがある場合には、これらレイアウトグループを構成するページファイルをレイアウト共有グループとして抽出する。

【0028】ページ間の距離を算出する手法はいくつか考えられる。ここではレイアウトタグおよびその特徴値に重み付けを行い、それぞれのタグ間の距離の総和をページ間距離とする手法を一例として説明する。距離を計算する2つのページに含まれるレイアウトタグの構造記述式の集合をそれぞれA、Bとするとページ間距離Dは以下のような式で表される。

$$【0029】D = \sum d_i(T_i)$$

ここで、 T_i はAUBを満たすレイアウトタグのi番目の要素であり、 d_i はレイアウトタグ T_i の距離関数である。ただし、 i は $1 \leq i \leq (A \cup B \text{ を満たすレイアウトタグの総数})$ である。

【0030】距離関数 d_i はレイアウトタグ T_i の関数

12

であり、 $T_i \in (A \cap B)$ の場合には、

$$d_i(T_i) = W_i * \sum W_{c_{ij}} * (f_i(C_{A_{ij}}), C_{B_{ij}}))$$

その他の場合には、

$$d_i(T_i) = W_i * L_i$$

である。ここで、 W_i はレイアウトタグ T_i の重み係数であり、たとえば「1」を例示できる。 C_{ij} はレイアウトタグ T_i における特徴値jの値である。 $W_{c_{ij}}$ はレイアウトタグ T_i における特徴値 C_{ij} の重み係数であり、たとえば「1」を例示できる。 f_i は特徴値間の距離を表す関数であり、たとえば特徴値が同一の時には「0」を返し、異なるときには「1」を返す関数を採用できる。また、 L_i はレイアウトタグ T_i が一方のページにのみ存在する場合の距離定数である。たとえば「 $L_i = 5$ 」を採用できる。

【0031】ページ間距離計算モジュール41は上記のような手法を用いてページ間距離Dを算出し、このページ間距離Dを利用してレイアウトグループ判別モジュール42が同一または類似のレイアウトをグループ化する。判別手段には、たとえばクラスタリング等の手法を用いることができ、類似の範囲の判断基準としてたとえば前記手法で算出されたレイアウト間距離Dが「10」以下のものを例示できる。

【0032】このようにして生成されたレイアウトグループを構成するページファイルの一例を図5に示す。図5(a)はあるページファイルをブラウジングして示した画面図であり、図5(b)は、他のページファイルをブラウジングして示した画面図である。前記した手法で求めたページ間距離はこの場合「0」である。すなわち、ページレイアウトの構造に関する限り図5(a)のファイルと図5(b)のファイルはレイアウトタグ、特徴値ともに同一である。よって、同一のレイアウトグループにグループ化される。ただし、レイアウト構造に関連しない内容(たとえばテーブル要素の各コンテンツ)が相違するのは勿論である。

【0033】また、同一レイアウトグループを構成するページファイルの他の例を図6に示す。図6(a)はあるページファイルをブラウジングして示した画面図であり、図6(b)は、他のページファイルをブラウジングして示した画面図である。前記した手法で求めたページ間距離はこの場合「3」である。両ページファイルはレイアウトタグの構造に関しては同一の構造をもつ。しかし、たとえば矢印で示した部分のレイアウトに関するレイアウトタグの特徴値(例示の場合表示色)が相違する。このような相違によりページ間距離が「3」と計算されている例である。ただし距離「10」以内なので類似と判断され両ページファイルは同一のレイアウトグループにグループ化される。

【0034】図7は、同一レイアウトグループにはグループ化されないページファイルの一例を示した画面図で

ある。(a)、(b)ともにページファイルがブラウジングされて示されているのは図5、6と同様である。この場合、両ページファイルのレイアウトタグの構造は同じである。しかし、特徴値の相違が大きく別レイアウトと判断された例である。たとえば矢印で示す部分のレイアウトタグ「td」は、図7(a)では「width」等を設定したうえで文字が配置されているのに対し、図7(b)では単に画像が配置されている。また、図7(a)ではレイアウトタグ「tr」において「bgcolor」が設定されているのに対し、図7(b)では設定されていない。これら相違によってページ間距離Dは「14」と計算され、異なるレイアウトグループにグループ化される。

【0035】以上のようにして同一または類似のページファイルがグループ化される。グループ化されたレイアウトグループは、データベース2に記録される。

【0036】次に、ページファイル内の一部領域のレイアウトが共通するレイアウト共有グループの抽出に係る部分について説明する。

【0037】代表値計算モジュール43は、前記手法で得られたレイアウトグループごとにレイアウトタグおよびその特徴値の代表値を導出する。まず、レイアウトグループを代表するレイアウトタグを導出する。代表タグの導出方法としては、レイアウトグループ内の各ページファイルに含まれるレイアウトタグの和集合や積集合を取る手法が例示できる。あるいは、あるレイアウトタグが存在するページファイル数がある閾値を超えたレイアウトタグについての集合を取る方法を例示できる。その他レイアウトグループのレイアウトタグを代表する任意の手法を用いることができる。次に、選択されたレイアウトタグについて、それぞれの特徴値を決定する。特徴値を決定するための手法としては、レイアウトグループ内の各ページファイルが持つ特徴値の多数決あるいは平均を取る方法などが例示できる。

【0038】レイアウト間距離計算モジュール44は、代表値計算モジュール43によって導出されたレイアウトグループごとの代表値を利用して、レイアウトグループ間の距離を計算する。

【0039】レイアウトグループ間の距離を算出する手法はいくつか考えられる。ここではレイアウトタグおよびその特徴値に重み付けを行い、それぞれのタグ間の距離の総和をレイアウト間距離とする手法を一例として説明する。距離を計算する2つのレイアウトグループの代表タグの集合をそれぞれA'、B'とするとレイアウト間距離D'は以下のような式で表される。

【0040】 $D' = \sum d_i'(T_i)$ ここで、 T_i はA' ∪ B'を満たすレイアウトタグのi番目の要素であり、 d_i' はレイアウトタグ T_i の距離関数である。ただし、 i は $1 \leq i \leq (A' \cup B')$ を満たすレイアウトタグの総数である。

【0041】距離関数 d_i' はレイアウトタグ T_i の関数であり、 $T_i \in (A' \cap B')$ の場合には、

$$d_i'(T_i) = W_i' * (M_i + \sum W_{c_{ij}}' * (f_{ij}'(C_{A_{ij}}), C_{B_{ij}})))$$

その他の場合には、

$$d_i'(T_i) = W_i' * L_i'$$

である。ここで、 W_i' はレイアウトタグ T_i の重み係数であり、たとえば「1」を例示できる。 C_{ij} はレイアウトタグ T_i における特徴値jの値である。

$W_{c_{ij}}'$ はレイアウトタグ T_i における特徴値 C_{ij} の重み係数であり、たとえば「1」を例示できる。

f_{ij}' は特徴値間の距離を表す関数であり、たとえば特徴値が同一の時には「0」を返し、異なるときには

「1」を返す関数を採用できる。 M_i は、レイアウトタグが双方のレイアウトグループに存在する時の距離定数である。 L_i' はレイアウトタグ T_i が一方のレイアウトグループにのみ存在する場合の距離定数である。このようにしてレイアウトグループ間の距離D'が計算できる。

【0042】レイアウト共有グループ判別モジュール45は、レイアウト間距離計算モジュール44によって導出されるレイアウト間距離D'を利用し、クラスタリングなどの手法を用いることによりグループ化を行う。そして、レイアウトの一部を共有していると思われるページ群(レイアウト共有グループ)の候補を列挙する。

【0043】なお、前記したレイアウトグループあるいはレイアウト共有グループには固有のレイアウトIDを割当てる。

【0044】アノテーション付加モジュール6は、ユーザからのアノテーション付加10に応答して、グループ単位でアノテーションを付加する。アノテーション付加モジュール6は、レイアウトグループに対して割当てられた固有のレイアウトIDとアノテーションを関連付けることにより、レイアウトグループ全体へのアノテーション付加を実現する。

【0045】アノテーションの付加に際しては、ページ群検出モジュール5により検出されたページ群(レイアウトグループあるいはレイアウト共有グループ)をユーザに提示する。この際、レイアウトの共有関係をグラフ表記などを用いて表す事により、ユーザが容易に理解できるように提示することができる。

【0046】提示されたページ群の中からユーザがアノテーションを付けるページを選択し、アノテーションを付加する。そうすると、当該ページのレイアウトIDに関連付ける形でアノテーションがデータベース2に保存される。この際、レイアウト共有グループが存在した場合には、共通して保持しているタグ構造(以下、共通レイアウトという)に付加されたアノテーションを複製し、各レイアウト共有グループのレイアウトIDに関連付けて保存する。

【0047】ユーザがアノテーション付けを行うページとして、既に共通レイアウト部分にアノテーション付けされたページを選択した場合には、共通レイアウト部分を強調表示するとともに、アノテーション情報を参照できるような形でユーザに提示することができる。従って、ユーザは当該レイアウトグループが独立して保持している部分のみにアノテーション付けを行うだけでページ全体へのアノテーション付加が行える。

【0048】距離計算式修正モジュール7は、ユーザがレイアウトグループの分割や統合を行ったり、あるいは共有関係の分離を行ったときに、そのような分割・統合あるいは分離を反映するように距離計算の各パラメータを修正する機能を持つ。

【0049】ユーザが提示されたページ群を統合・分割するなどの修正を加えた場合には、その結果を用いてページ間距離計算式を修正し、以降のページ群分割の精度を向上させることができる。このような修正手法には各種の手法が考え得る。ここでは、レイアウトタグおよび特徴値への重み付けの変更によりページ間距離式を変更する手法について述べる。

【0050】レイアウトグループの分割が指示された場合は、分割後のグループ間で相違のあるレイアウトタグおよび特徴値に注目し、その相違するレイアウトタグおよびその特徴値の重みを増加することにより、以降のページ群検出において別のレイアウトグループとして検出されるようにページ間距離計算式を変更する。なお、分割後のグループ間で一致するレイアウトタグおよび特徴値の重みを減少させても良い。

【0051】レイアウトグループの合併（統合）が指示された場合は、上記とは逆に、相違のあるレイアウトタグおよび特徴値の重みを減少することにより、以後のページ群検出およびレイアウト共有判別において同一ページ群もしくはレイアウト共有グループと判別されるよう式の変更を行う。なお、併合後のグループ間で一致するレイアウトタグおよび特徴値の重みを増加させても良い。

【0052】また、レイアウト共有関係の解除（分離）などの修正を加えた場合には、上記と同様にしてレイアウトグループの代表値間で相違のあるレイアウトタグおよび特徴値に注目し、その重み付け等を変更することによりレイアウト間距離計算式を修正し、以降のレイアウト共有判別の精度を向上させることができる。

【0053】以上、本実施の形態の情報処理システムの全体を説明した。次に、このシステムを用いてアノテーション付加を行う方法について説明する。

【0054】まず、ユーザは目的のサイトのURLとアノテーションをつける対象の条件（ディレクトリや更新日時など）を指定する。そうすると、前記システムはページ取得モジュール3により対象となるHTMLファイルを取得し、HTMLファイル解析モジュール4による

ページファイルの解析の後、ページ群検出モジュール5によりレイアウトグループおよびレイアウト共有グループの検出を行う。

【0055】次に、たとえば構成するページファイルの数の多い順等、任意の順番でユーザにレイアウトが同一であると推定されたページ群（レイアウトグループ）を提示し、ページ群内の任意のページ（ページファイル）にアノテーション付加を要求する。

【0056】図8は、アノテーション付加の処理の一例を示したフローチャートである。まず、前記のとおり、データベース2からレイアウトグループ（ページ群）を取得し、ユーザに提示する（ステップ50）。その後、全レイアウトグループについてアノテーションが付加されたかを判断し（ステップ51）、全レイアウトグループについてアノテーションが付加されている場合には処理を終了する（ステップ52）。アノテーションが付加されていないレイアウトグループが残存する場合にはステップ53以降の処理に進む。

【0057】ステップ53では、まず任意のレイアウトグループ（ページ群）を選択する（ステップ53）。このページ群に関連付けるレイアウトIDとして、レイアウトID（1）が決定される（ステップ53）。

【0058】次に、ユーザによりページ群（レイアウトグループ）内の任意のページ（ページファイル）が選択される（ステップ54）。選択されたページファイルは適当なブラウザによってユーザに提示される。ユーザはこの表示画面を見ながらアノテーションを付加する（ステップ55）。たとえばPDAや小画面デバイスのための画面分割や音声ブラウザのためのコンテンツ内容ごとへのジャンプ用リンクを付加する。この付加されたアノテーションにはレイアウトID（1）が付与され、関連付けられる（ステップ55）。

【0059】アノテーションの付与後、たとえばページ群内の適用可能ページ数を提示し、このページ全体に付与したアノテーションを提供するか、あるいは個別にアノテーション付与の適用を判断するかをユーザに選択させる。つまり、ページ群（レイアウトグループ）全体へのアノテーションの適用が可能かを判断する（ステップ56）。ステップ56の判断がyesの場合、ページ内の全ページファイルにレイアウトID（1）が付与される（ステップ57）。その後、レイアウト共有グループへのアノテーション付与のステップ（ステップ58）に進む。

【0060】一方、ステップ56の判断がnoの場合、ページ群内の各ページについて前記付与されたアノテーションの適用が可能かを判断する必要がある。ステップ59においてページ群の全ページについて確認が終了したかを判断し（ステップ59）、この判断がnoの場合にはページ群内の残るページの何れかを選択する（ステップ60）。

10

20

30

40

50

【0061】選択されたページについて、アノテーションの適用が可能かを判断し（ステップ61）、適用可能な場合（ステップ61の判断がy e sの場合）には選択ページにレイアウトID（1）を付与する（ステップ62）。適用不可の場合（ステップ61の判断がn oの場合）には選択ページに仮レイアウトIDを付与する（ステップ63）。この仮レイアウトIDはレイアウトID（1）が適用できないものについて付与する共通のIDであり、後に説明するように個別の処理を行うための識別IDである。

【0062】レイアウトID（1）か仮レイアウトIDの何れかが付与された後、ステップ59に戻り、ステップ59以下の処理を行う。

【0063】一方、ステップ59でページ群内の全ページの確認が終了したと判断された場合、仮レイアウトIDが付与されているページが存在するかを判断し（ステップ64）、存在する場合（y e sの場合）には仮レイアウトIDを持つページ群へのアノテーション付与の処理（ステップ65）に進む。仮レイアウトIDを持つページがないときにはステップ58に進む。

【0064】図9は、仮レイアウトIDが付与されたページ群にアノテーションを付与する処理の一例を示したフローチャートである。図8のフローチャートにおいてステップ65に進んできた場合、この処理が行われる。

【0065】まず、仮レイアウトが付与されているページで構成されるページ群内の任意のページを選択する（ステップ70）。この選択により選択ページにレイアウトID（2）が付与される。次に、選択ページに前記と同様にアノテーションを付加する（ステップ71）。アノテーションにはレイアウトID（2）が付与される。

【0066】次に、仮レイアウトIDのページ群全体に付与されたアノテーションが適用できるかを判断する（ステップ72）。この判断がy e sの場合には、仮レイアウトIDのページ群の全ページにレイアウトID（2）を付与し（ステップ73）、ページ間距離計算式の修正を行って（ステップ74）、処理を終了する（ステップ75）。

【0067】ステップ72での判断がn oの時（仮レイアウトIDのページ群全体にアノテーションが適用できないとき）は、個々のページについてアノテーションが適用できるかを判断する必要がある。この場合ステップ76において仮レイアウトIDページ群内の全ページについて確認が終了したかを判断し（ステップ76）、終了していない場合（判断がn oの場合）は、ページ群内の任意のページを選択し（ステップ77）、選択ページへのアノテーション適用が可能かを判断し（ステップ78）、適用可能な場合は選択ページにレイアウトID（2）を付与（ステップ79）した上でステップ76に戻る。適用できない場合はそのまま（仮レイアウトID

を保持したまま）ステップ76に戻る。

【0068】一方、ステップ76での判断がy e sの場合（全ページの確認終了後）、仮レイアウトIDが付与されているページがあるかを判断し（ステップ80）、存在しない場合（判断がn o）はステップ74に進んでページ間距離計算式の修正を行った上で処理を終了し（ステップ75）、存在する場合（ステップ80の判断がy e sのとき）にはステップ70に戻って前記の処理を繰り返す。

10 【0069】このようにして、仮レイアウトIDの付与されたページがなくなり、着目しているページ群（レイアウトグループ）の全てについて妥当なアノテーションが付与されることになる。なお、同一レイアウトグループ内で異なるアノテーションが付与されたときにはステップ74でページ間距離計算の修正が行われるので、次回以降のページ間距離の計算ではこの結果が反映されて異なるレイアウトグループにグループ化されるよう学習することになる。

20 【0070】次に、レイアウト共有グループへのアノテーション付加の処理（ステップ58）を説明する。図10は、レイアウト共有グループへのアノテーション付加の処理の一例を示したフローチャートである。まず、レイアウト共有グループ内の任意のページ群（レイアウトグループ）を選択する（ステップ81）。

30 【0071】次に、共有レイアウトへのアノテーションに複数候補があるかを判断する（ステップ82）。ページ群の分割やレイアウト共有グループ内で異なるアノテーションが付加されることなどにより、レイアウト共有部分に複数のアノテーションの候補が存在することが考えられる。この様な場合には、以降のレイアウト共有グループにアノテーションを付加する際には、アノテーションの候補を順に提示し、どのアノテーションを付加するか選択させる（ステップ83）。

40 【0072】次に、共有レイアウト部分に提示されたアノテーションが適用可能かを判断し（ステップ84）、適用可能な場合には共有部分へのアノテーションの複製を行い、さらに共有部分以外の部分についてアノテーションの付与を要求する（ステップ86）。アノテーション付与の方法は前記したとおりである。このように共有部分については予め付与されているアノテーションを複製することが可能になり、ユーザは共有部分以外の部分についてだけアノテーションを付与すればよい。これによりアノテーション付与に係る作業負担を軽減することができる。一方、共有部分へのアノテーションの適用が不可の場合、ページ全体へのアノテーションの付与を要求する（ステップ85）。その後、仮レイアウトIDを持つページ群内でもアノテーション付加と同様の処理を行い（ステップ87）、レイアウト共有グループ内の全てのページ群について前記処理が行われたかを判断し（ステップ88）、処理が終了している場合は終了し

(ステップ89)、未処理の場合はステップ81に戻って前記処理を繰り返す。なお、共有レイアウトの全体にアノテーションの適用がなされない場合には、前記同様レイアウト間距離の計算式の修正が行われることになる(ステップ87)。

【0073】以上図8～10に示した処理を全てのページ群に対して順に行うことにより、サイトへのアノテーションの一斉付加が終了する。

【0074】以上説明したように本実施の形態の情報処理システムあるいは方法によって、レイアウトが同一あるいは類似するページについては一斉にアノテーションを付与・適用することが可能になる。また、ページの一部に共有レイアウトを有する場合には、この共有部分へのアノテーション付加・適用も簡略化される。これによりユーザによるアノテーション付与の作業が大幅に効率化される。特にニュースサイトやデータベースサイトのように大量のページファイルを有し、これらページのレイアウトに共通する部分が多いものについてはその作業性の向上は著しい。

【0075】また、システムによって自動的に判定された類似の判断がユーザによって変更された場合には、前記のとおり距離計算式の修正を行うことによりシステムによって自動的に判定基準が修正される。これにより、グループ化の精度を向上できる。また、この判定基準の変更はユーザのアノテーション付与の作業に伴って行われるものであり、ユーザは単にアノテーション付与作業を行うだけで自動的に作業性が向上するメリットを享受する。すなわち、レイアウトグループあるいはレイアウト共有グループの判定基準はユーザ作業によって自動的に向上する学習効果を有する。

【0076】なお、前記実施の形態では、アノテーションの一斉付与の一例を説明したが、既に付与されたアノテーションを利用して以下のようなページファイルへの動的アノテーション付与とトランスコーディングを行うこともできる。

【0077】すなわち、ユーザがHTML文書を読覧中に、特定の位置へマーキングなどのアノテーション付加を行う。システムではこの情報を当該ページのレイアウト情報(レイアウトタグ・特徴値)とともに保存しておき、以降の閲覧時には、この情報を利用して画面の分割やマーキングされた位置へのリンクを埋める等のトランスコーディングを行うことができる。

【0078】また、アノテーションが付加されていないページの閲覧が要求された場合は、既にアノテーションが登録されたページと要求ページとの間でページ間距離計算モジュールにより距離計算を行う。この結果、ページ間距離が、ある閾値よりも小さい場合には、最も近いページのアノテーションを用いてトランスコーディングを行いユーザに提示する。ここで、ユーザからアノテーションの誤りを指摘された場合には 距離計算式修正モ

ジュールにより、距離計算式の変更を行う。ユーザは必要があれば、新たにアノテーション情報を付加する。この手法により、あらかじめ全てのページに対してアノテーションを付加するのではなく、ユーザが閲覧時に必要に応じてアノテーションを付加していくことにより、段階的にサイト全体へのアノテーション付加が可能となる。

【0079】以上、本発明者によってなされた発明を発明の実施の形態に基づき具体的に説明したが、本発明は前記実施の形態に限定されるものではなく、その要旨を逸脱しない範囲で種々変更可能であることは言うまでもない。

【0080】たとえば、前記実施の形態では、ページファイル間の類似性の判断手法にページ間距離あるいはレイアウトグループ間の距離をレイアウトタグと特徴値に重み付けして算出する手法を例示した。しかし、これに限られず、たとえばタグスケルトンの手法を用いたり、画像あるいはHTML文書の内容(テキスト)の類似性を基準に判断しても良い。

【0081】また、前記実施の形態における、レイアウト共有グループの導出およびレイアウト共有グループを用いた共有レイアウトへのアノテーションの適用について、これを本発明の必須の要件にする必要はない。つまり、レイアウトグループの導出とレイアウトグループ内へのアノテーションの適用に限る場合も本発明とすることができる。この場合であっても、アノテーション付与の作業負担の軽減という本発明の効果を十分に達成することができる。また、同様に、ページ間あるいはレイアウトグループ間の距離計算式の修正に関する要件も本発明の必須の要件にする必要はない。同様に本発明の効果を達成は十分に可能である。

【0082】また、前記実施の形態では、HTML文書のレイアウトの類似性に着目してグループ化することを説明したが、レイアウトに関係しないタグや、文書の内容についての類似性の判断に本実施の形態を拡張することができる。この場合、HTML文書の構造あるいは内容自体の類似性を判断することになり、たとえばサイト管理者用にサイト内解析に用いたり、サイト内のページファイル変更履歴解析に用いることができる。

【0083】また、前記実施の形態では、ページファイルとしてHTMLファイルを例示したが、XML(Extensible Markup Language)やダイナミックHTML等他のマークアップ言語を用いて記述されているページファイルにも適用できることは勿論である。

【0084】

【発明の効果】本願で開示される発明のうち、代表的なものによって得られる効果は、以下の通りである。すなわち、本発明によって、ページファイルへのアノテーション付与の作業を効率的に行うことができる。また、本発明のシステムの使用に従ってより高精度なレイアウト

グループまたはレイアウト共有グループのグループ化を実現できる。

【図面の簡単な説明】

【図 1】本発明の一実施の形態である情報処理システムの一例を示したブロック図である。

【図 2】HTMLファイル解析モジュールの構成の一例を示したブロック図である。

【図 3】URLとそのURLに関連するレイアウトタグと特徴値の例を示した図である。

【図 4】ページ群検出モジュールの構成の一例を示した 10
ブロック図である。

【図 5】同一レイアウトグループにグループ化されるページファイルの一例をブラウジングして示した画面図である。

【図 6】同一レイアウトグループにグループ化されるページファイルの他の例をブラウジングして示した画面図である。

【図 7】同一レイアウトグループにはグループ化されないページファイルの一例をブラウジングして示した画面図である。

20

【図 8】アノテーション付加の処理の一例を示したフローチャートである。

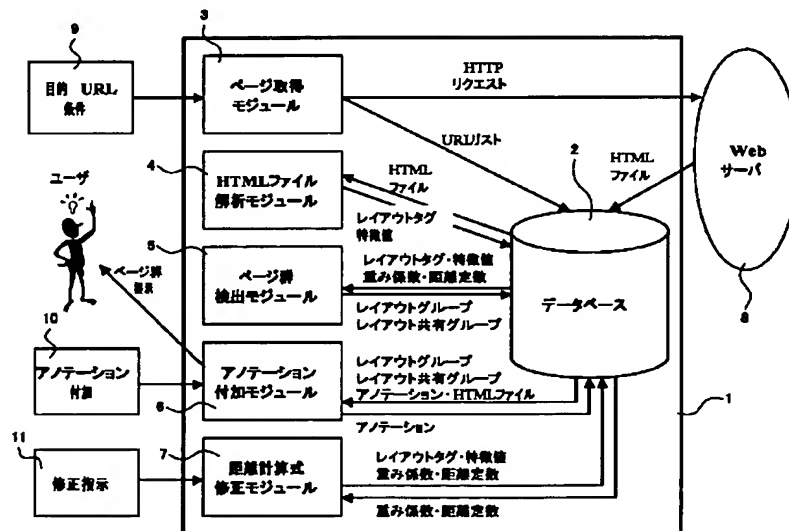
【図 9】仮レイアウトIDが付与されたページ群にアノテーションを付与する処理の一例を示したフローチャートである。

【図 10】レイアウト共有グループへのアノテーション付加の処理の一例を示したフローチャートである。

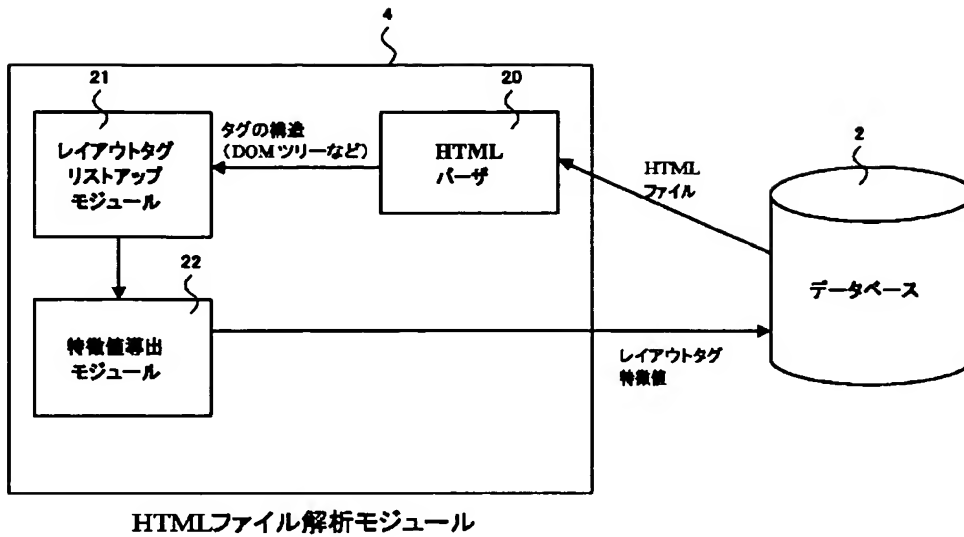
【符号の説明】

1…情報処理システム、2…データベース、3…ページ取得モジュール、4…HTMLファイル解析モジュール、5…ページ群検出モジュール、6…アノテーション付加モジュール、7…距離計算式修正モジュール、8…ウェブサーバ、9…目的URL条件、10…アノテーション付加、20…HTMLパーザ、21…レイアウトタグリストアップモジュール、22…特徴値導出モジュール、41…ページ間距離計算モジュール、42…レイアウトグループ判別モジュール、43…代表値計算モジュール、44…レイアウト間距離計算モジュール、45…レイアウト共有グループ判別モジュール。

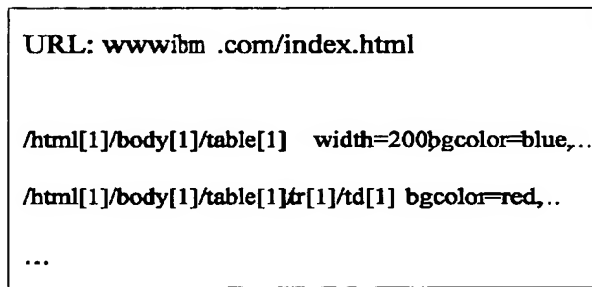
【図 1】



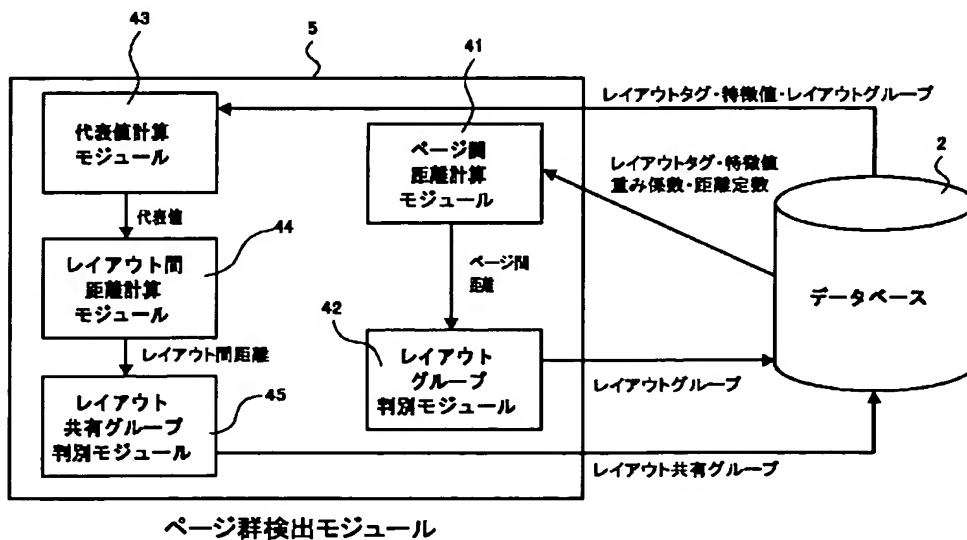
【図 2】



【図 3】

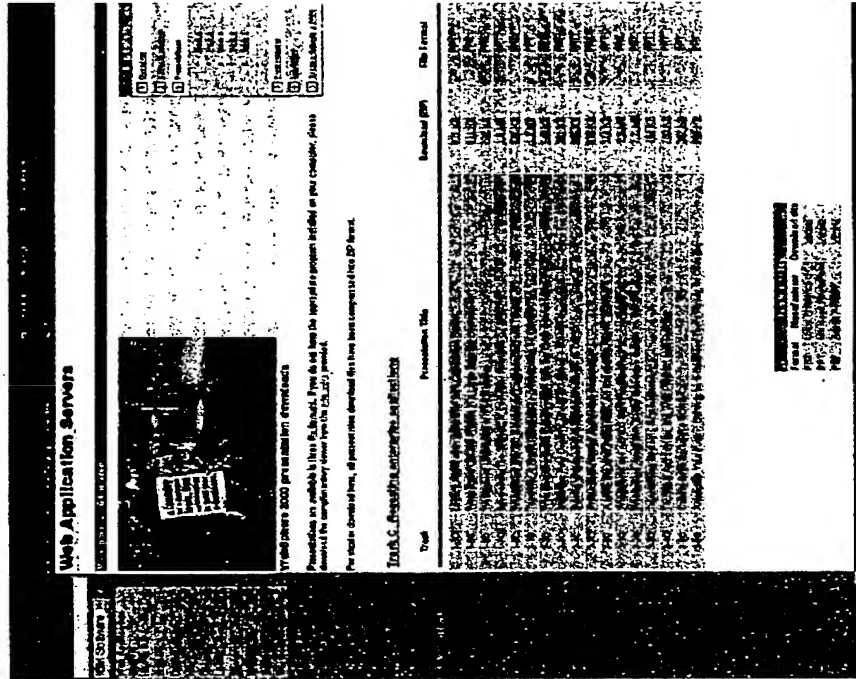


【図 4】

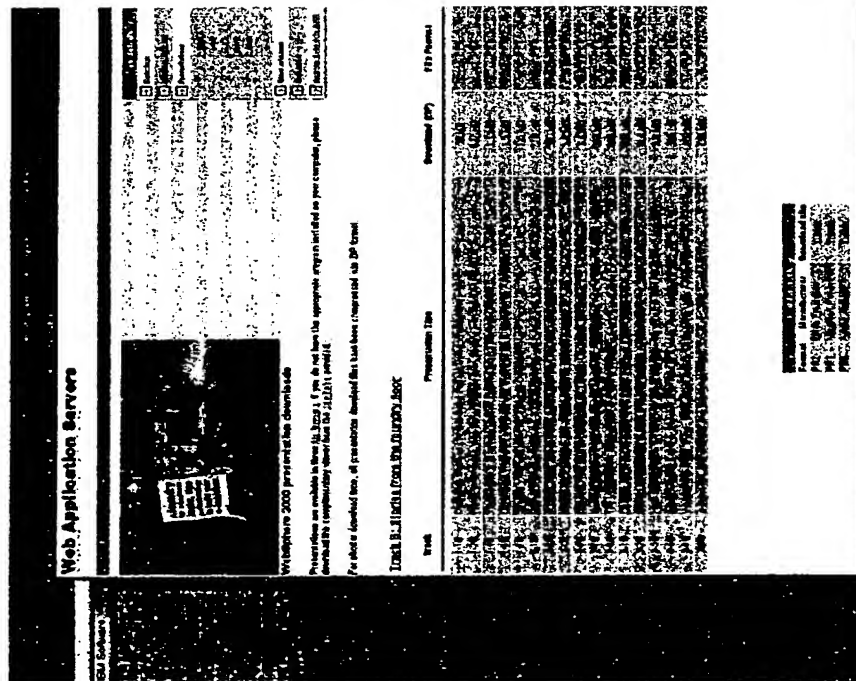


【 図 5 】

(b)

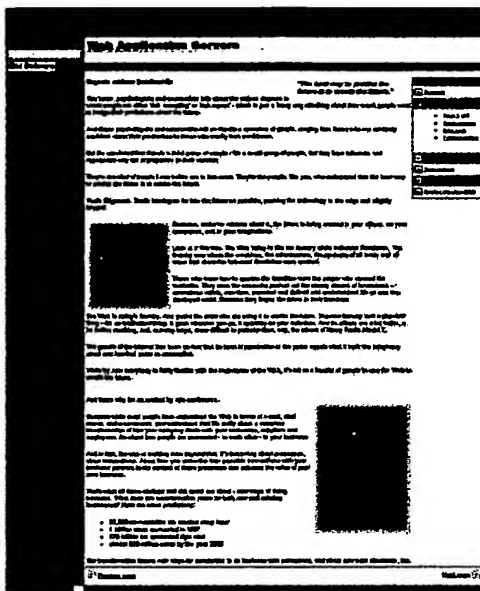


(a)

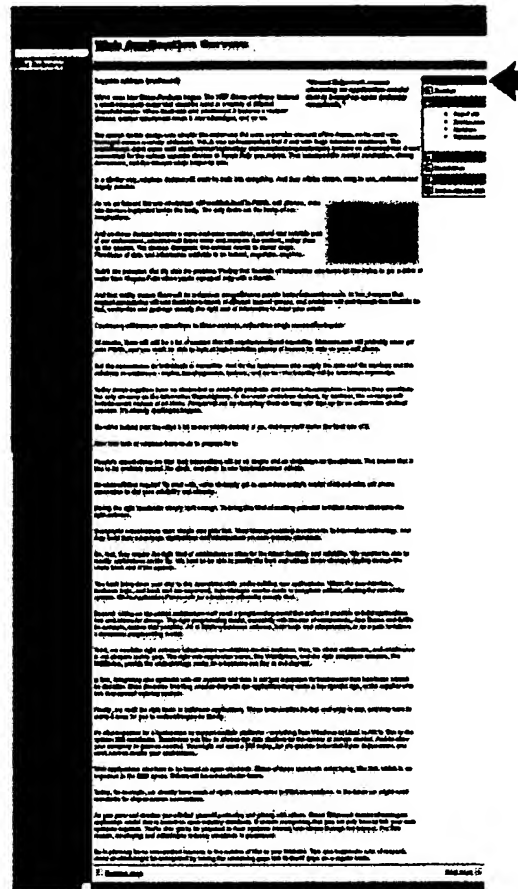


【図6】

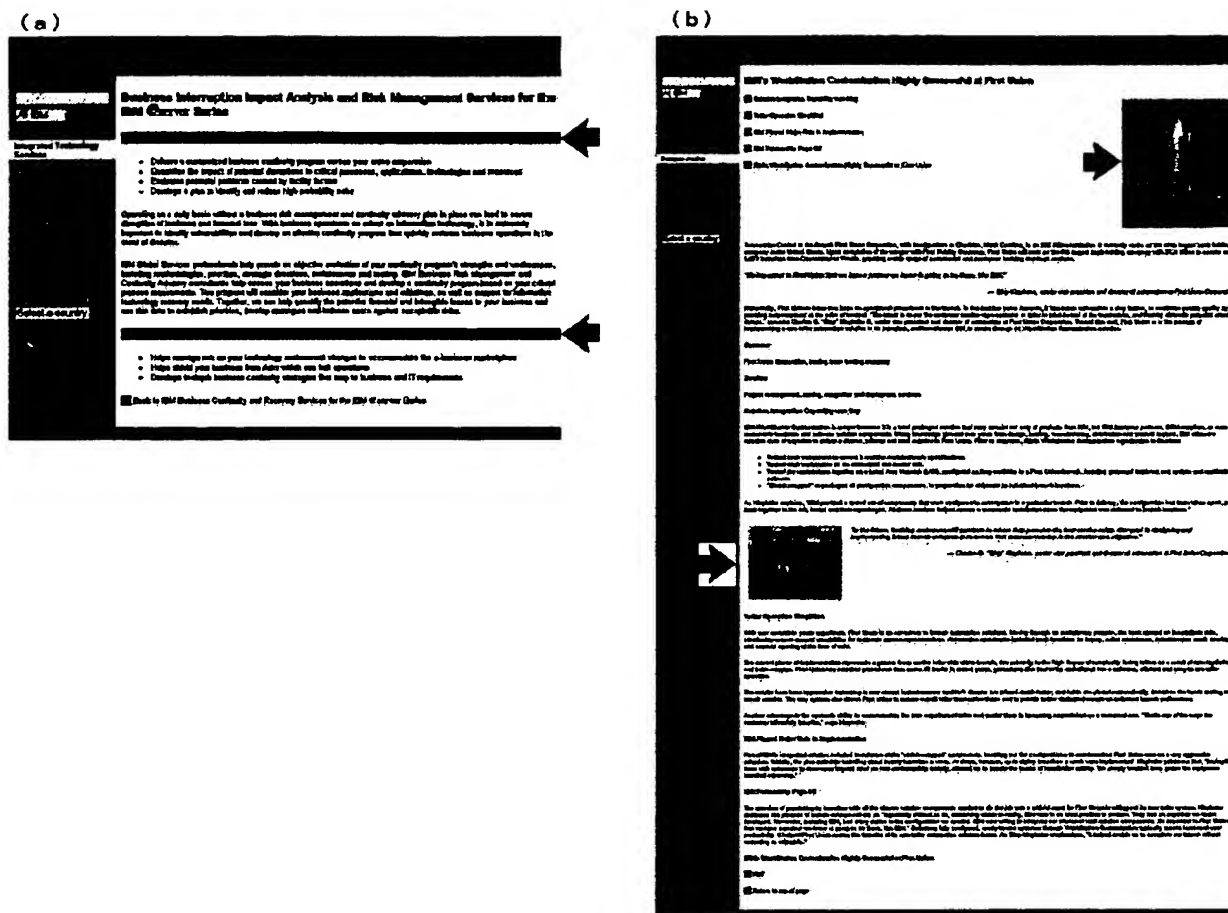
(a)



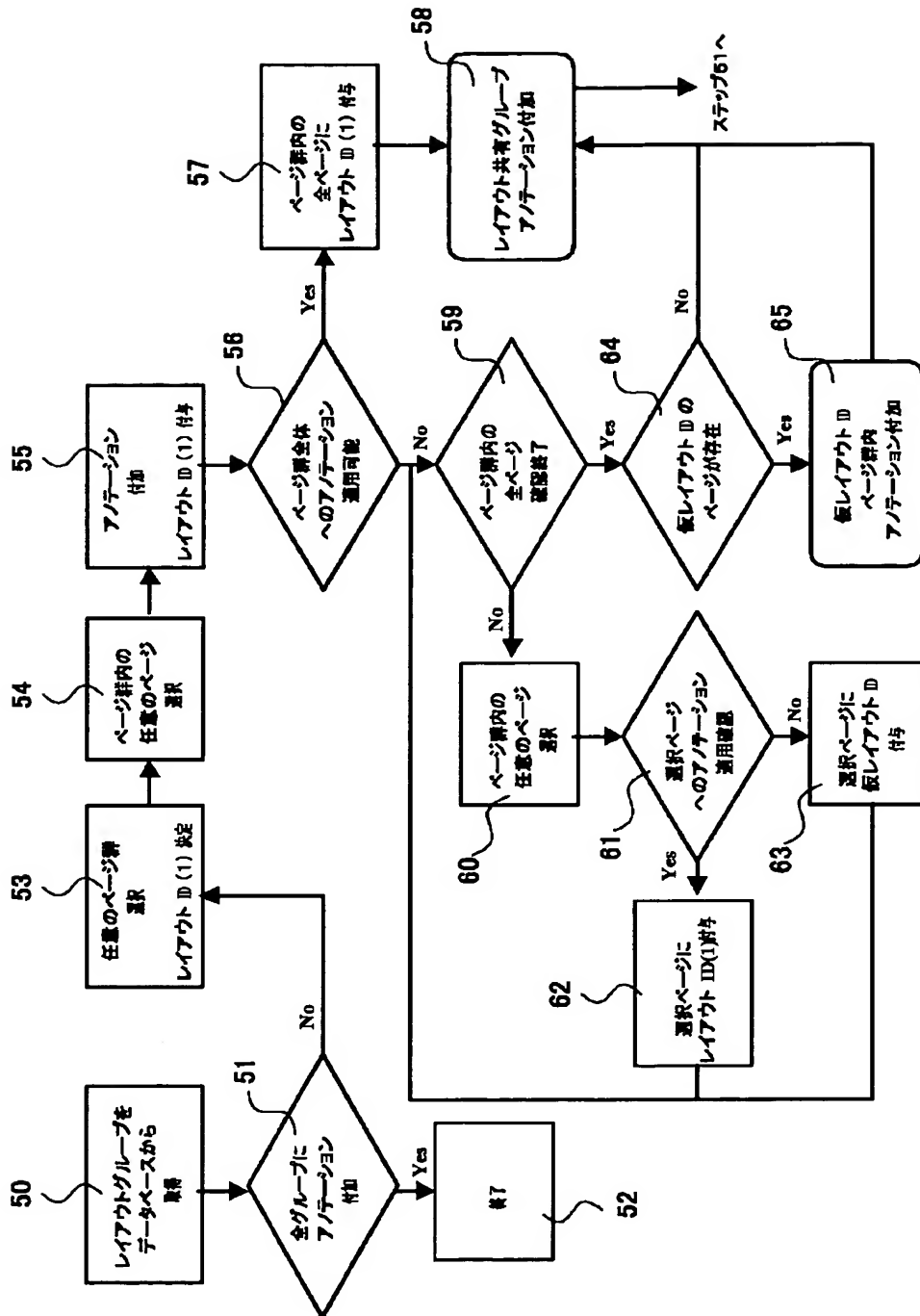
(b)



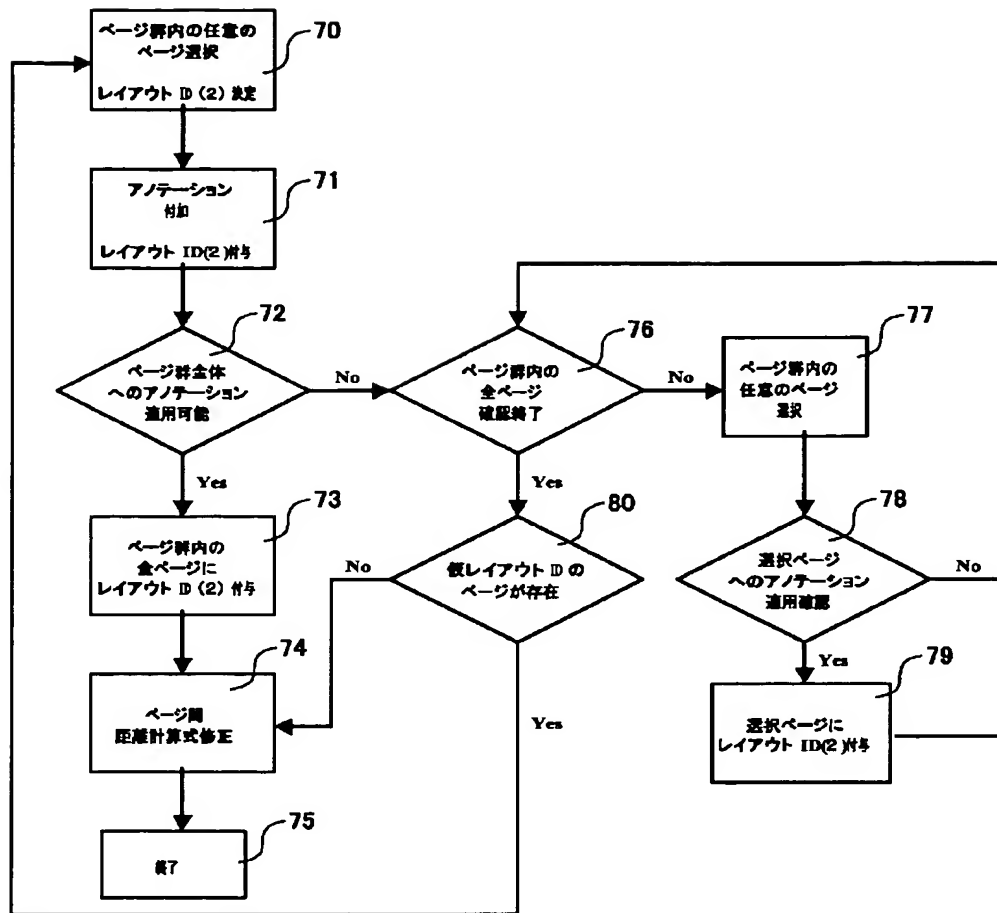
【图7】



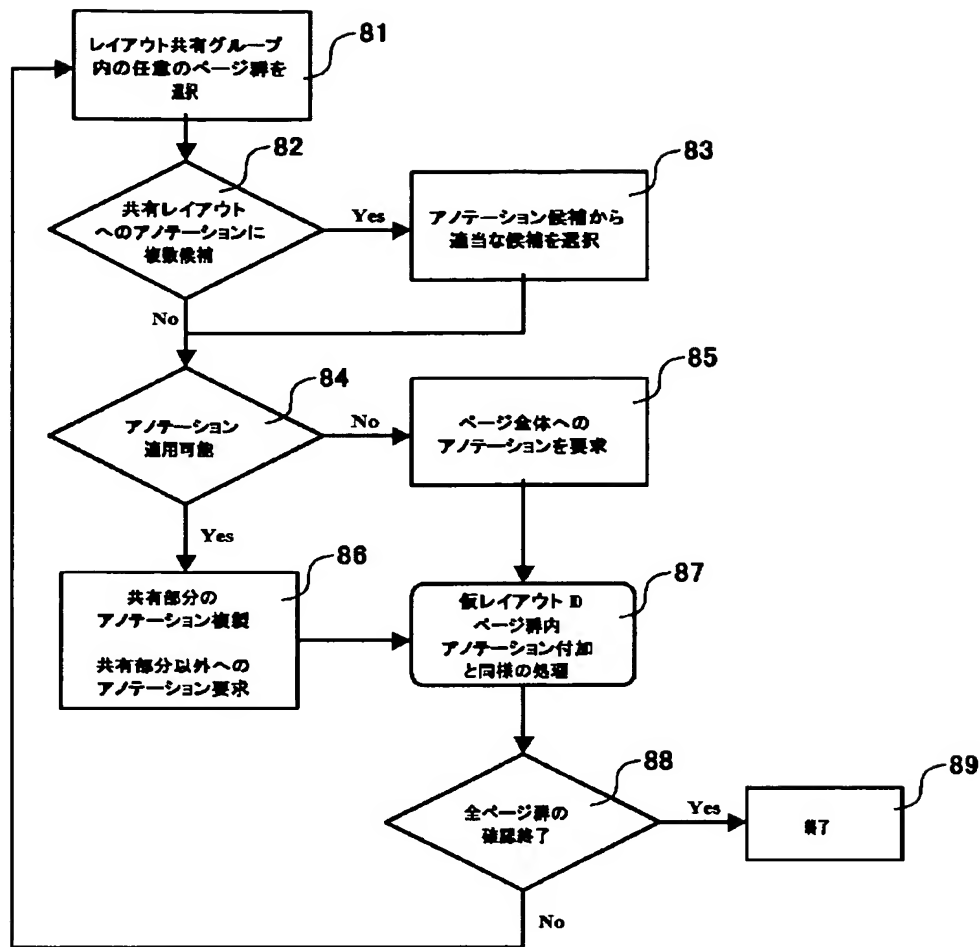
【图8】



【図9】



【図10】



フロントページの続き

(72)発明者 福田 健太郎
神奈川県大和市下鶴間1623番地14 日本ア
イ・ビー・エム株式会社 東京基礎研究所
内
(72)発明者 ▲高▼木 啓伸
神奈川県大和市下鶴間1623番地14 日本ア
イ・ビー・エム株式会社 東京基礎研究所
内

(72)発明者 前田 潤治
神奈川県大和市下鶴間1623番地14 日本ア
イ・ビー・エム株式会社 東京基礎研究所
内
(72)発明者 浅川 智恵子
神奈川県大和市下鶴間1623番地14 日本ア
イ・ビー・エム株式会社 東京基礎研究所
内

Fターム(参考) 5B069 AA01 BA03 BB16 CA06 FA01
5B075 ND36 NR12 NS10 PP02 PP03
PQ02 PQ42 PR06 UU06

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☒ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.